Journal of Computing Research and Innovation (JCRINN) Vol. 5 No. 3 (2020)
eISSN: 2600-8793
10.24191/jcrinn.v5i3.149

Forecasting Air Pollution Index (API) $PM_{2.5}$ Using Support Vector Machine (SVM)

Nor Hayati Shafii[1]*, Prof Madya Rohana Alias[2], Nur Fithrinnissaa Zamani[3], Dr Nur Fatihah Fauzi[4]

[1,2,4] Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia

*Corresponding author: *norhayatishafii@uitm.edu.my*
Received Date: 28 April 2020
Accepted Date: 16 May 2020
Revised Date: 30 May 2020
Published Date: 1 June 2020

ABSTRACT

Air pollution is a current monitored problem in areas with high population density such as big cities. Many regions in Malaysia are facing extreme air quality issues. This situation is caused by several factors such as human behavior, environmental awareness and technological development. Accessing the air pollution index accurately is very important to control its impact on environmental and human health. The work presented here aims to access Air Pollution Index (API) of $PM_{2.5}$ accurately using Support Vector Machine (SVM) and to compare the accuracy of four different types of the kernel function in Support Vector Machine (SVM). SVM is relatively memory efficient and works relatively well in high dimensional spaces data which is better than the conventional method. The data used in this study is provided by the Department of Environment (DOE) and it is recorded from two Continuous Air Quality Monitoring Stations (CAQM) located at Tanah Merah and Kota Bharu. The results are analyzed using mean absolute error (MAE) and root mean squared error (RMSE). It is found that the proposed model using Radial Basis Function (RBF) with its parameters of cost and gamma equal to 100 can effectively and accurately forecast the API based on the model testing with 0.03868583 (MAE) and 0.06251793 (RMSE) for API in Kota Bharu and 0.03857308 (MAE) and 0.05895648 (RMSE) for API in Tanah Merah.

Keywords: *API, Support Vector Machine (SVM), time series forecasting, kernel function, $PM_{2.5}$*

INTRODUCTION

Air pollution can be defined as any substance that can harm human health and other living species. It will affect the different organs and systems in human such as the respiratory system, cardiovascular system, nervous system, urinary system, digestive system as well as harming pregnancies. Based on Kampa and Castanas (2008), air pollutants can be divided into four categories which are gaseous pollutants, persistent organic pollutants, heavy metals and particulate matters. The issues of air pollution have commonly become a big problem in Malaysia. Monitoring air pollution levels is very important to detect pollution peaks, to improve the air pollution control and eventually the air quality.

From previous studies, there are several methods used to access air pollution index such as autoregressive integrated moving average (ARIMA), fuzzy time series (FTS), artificial neural network (ANN), support vector machine (SVM) etc. The study from Vinagre et al. (2016) shows SVM gives good forecasting on

energy consumption compared to their previous study using the same data with the ANN method. Another study was conducted by Vaiz and Ramaswami (2016) for predicting the stock trend by combining two methods of SVM and ANN gives the best accuracy of forecasting rather than just predict using the ANN. Oloruntoba and Akinode (2017) used all possible data mining techniques such as linear regression, SVM, Decision Tree, Lasso Regression (LASSO), ElasticNet (EN) and K-Nearest Neighbour (KNN) to predict student performance and finally by comparing the value of MSE, the results indicated that SVM was the best method compared to the others. It was done by tuning the parameters of the SVM algorithm i.e. kernel to improve the accuracy of forecasting and lower the MSE value. Last but not least, the study conducted by Arampongsanuwat and Meesad (2011) using Support Vector Regression (SVR) model with Gaussian Radial Basis Kernel functions to forecast PM_{10} in Bangkok also produced best result with lowest MSE. The study reported that the model was satisfactory and the technique of SVR can be used to predict $PM_{10.}$ As a conclusion, SVM model provides promising alternatives and advantages in time series forecasting as SVM model provides a few of free parameters compared to other conventional neural network models, gives a better prediction than the conventional model due to the adoption of the structure Risk Minimization Principle and also can abolish the typical drawbacks of conventional neural network model such as overfitting training and local minima, and evidence to be more expandable and stronger.

Therefore, this study focuses on SVM with four different algorithms i.e. kernels to access API and comparing the accuracy performance of these four kernels in predicting the API index based on the lowest value of mean absolute error (MAE) and root mean squared error (RMSE). By predicting accurate future air pollution readings, all parties including government and society can take early precautionary action to preserve cleaner air before it is too late.

METHODOLOGY

Method of Data Collection

The data of Air Pollution Index (API) $PM_{2.5}$ used in this study are secondary data obtained from Malaysia's Open Data Portal (MAMPU, 2019) for Kelantan state from January to December 2018. Some of the data are also provided by the Department of Environment (DOE) and it was recorded from two Continuous Air Quality Monitoring Stations (CAQM) located in both industrial and urban areas at Tanah Merah and Kota Bharu.

Data Analysis

The process of data analysis is done by using R programming. The flow of the data analysis procedures is shown in Figure 1,

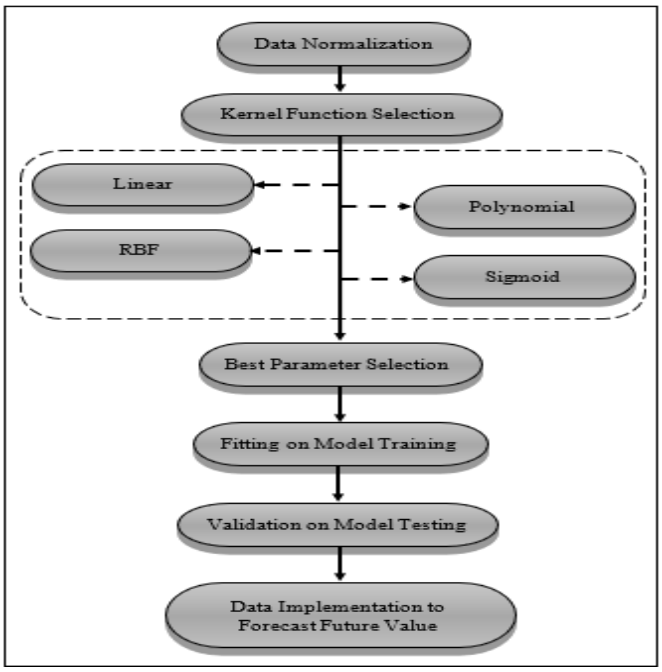

Figure 1: Flow of Data Analysis

I) Data Normalization

Preprocessing of the input data is scaling the data in the range [0,1] and checking for possible outliers. The scope of the data has been adjusted in the range [0,1] by using min-max normalization that leads to stable and accurate data in the forecasting results. Figure 2 represents the graph of the normalized air pollution index in Kota Bharu.

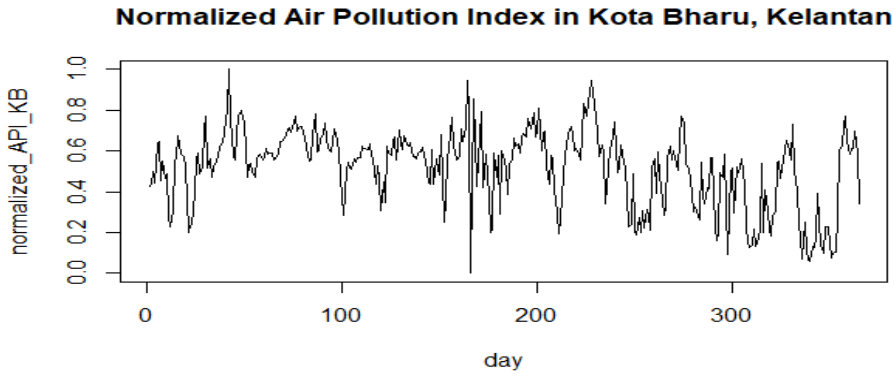

Figure 2: Normalized of API in Kota Bharu, Kelantan

After eliminating the outliers, the data is divided into two sections which are 70% for the training set and 30% for the testing set as shown in Table 1.

.

Table 1: Training and Testing set

API	Training set data (day)	Testing set data (day)

Tanah Merah	1-241	242-344
Kota Bharu	1-251	252-359

II) Model Development

i. Support Vector Machine (SVM)

SVM is applied in machine learning applications developed by Vapnik (1995); Vapnik et al. (1997) as cited in Lu and Wang (2005). The main idea of this method is to map the original data x into a feature space F with the higher dimensionality via non-linear mapping function ϕ which is, in general, is unknown and then carry on linear regression in the feature space.

Hence, the problem of estimating a function, the regression approximation has been addressed according to a given data set (where x_i as input vectors, d_i as desired values, y as target value), that is produced from the ϕ function. SVM method approximates the function as given in Equation (1),

$$y = \sum_{i=1}^{p} w_i \phi_i(x) + b = w\phi(x) + b \qquad (1)$$

where $w = [w_1, ..., w_p]$ are the weights vector, the bias coefficients, b and $\phi(x) = [\phi_1(x), ..., \phi_p(x)]$ represent the basis function vector. The regularized risk function, $R(C)$ is as given in Equation (2) with the error function defined by the ε-insensitive loss function, $L_\varepsilon(d_i, y_i(x))$ as given in Equation (3),

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i(x)) + \frac{1}{2}\|w\|^2 \qquad (2)$$

where,

$$L_\varepsilon(d_i, y_i(x)) = \begin{cases} |d_i - y_i(x)| - \varepsilon, \\ 0, \end{cases}$$
$$|d_i - y_i(x)| \geq \varepsilon$$
$$otherwise \qquad (3)$$

the term $\frac{1}{2}\|w\|^2$ is used for measuring the function's flatness, C is the regularized constant parameter that determines the trade-off between the training error and the model flatness. Equation (4) is necessary to minimize the regularized risk function by introducing the slack variables, ζ, ζ^* which are lead to Equation (2) with the constraint of Equation (5),

Minimize:

$$R(w, \zeta^*) = \frac{1}{2}\|w\|^2 + C^*\sum_{i=1}^{p}(\zeta_i + \zeta_i^*) \qquad (4)$$

subjected to:

$$\begin{aligned} w\phi(x_i)+b-d_i &\le \varepsilon+\zeta_i \\ d_i-w\phi(x_i)-b_i &\le \varepsilon+\zeta_i \\ \zeta,\zeta^* &\ge 0 \end{aligned} \tag{5}$$

Thus, the explicit form of Equation (1),

$$f(x,\alpha,\alpha^*)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)k(x_i,x_j)+b \tag{6}$$

In Equation (6) α_i and α^* are the Lagrange multipliers, which satisfy the following equalities,

$$\begin{aligned} \alpha_i*\alpha_i^* &= 0, \\ \alpha_i &\ge 0, \\ \alpha_i^* &\ge 0 \end{aligned}$$

where $i=1,\ldots,l,$ and can be obtained by maximizing the dual form of Equation (4),

$$\begin{aligned} \phi(\alpha,\alpha^*) &= \sum_{i=1}^{l} d_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i=1}^{l}(\alpha_i-\alpha_i^*) \\ &\quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(\alpha_i,\alpha_j) \end{aligned} \tag{7}$$

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0$$

with constrains: (8)

$$\begin{aligned} 0\le\alpha_i &\le C \qquad i=1,\ldots,l \\ 0\le\alpha_i^* &\le C \qquad i=1,\ldots,l \end{aligned}$$

The parameters which are epsilon (ε), cost (C) and gamma (γ) was set as a default $\left(\varepsilon=0.1, C=1, \gamma=\frac{1}{\text{data dimension}}\right)$. Based on the nature of quadratic programming, only those data corresponding to non-zero (α_i,α_i^*) pairs can be referred to as support vectors. $k(x_i,x_j)$ represents kernel function and obtained by $k(x_i,x_j)=\phi(x_i)*\phi(x_j)$ in the feature space, F. Hence, all the computations related to ϕ will be carried on by the kernel function in feature space.

ii. Kernel Function Selection

There are four types of kernel functions used to build the SVM model in forecasting the air pollution index; Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid. Generally, these kernel functions are given by Equation (9) to (12), where $x_i, x_j \in R^p$.

$$linear: k\left(x_i,x_j\right)=x_i\cdot x_j, \tag{9}$$

$$polynomial : k(x_i, x_j) = (x_i \cdot x_j + 1)^d, \quad (10)$$

$$RBF : k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right), \quad (11)$$

$$sigmoid : k(x_i, x_j) = \tanh(x_i \cdot x_j + 1), \quad (12)$$

The accuracy of these kernel functions is effectively compared based on the measurement of error. The error measures used are Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE),

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - Q_i|, \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - Q_i)^2}, \quad (14)$$

Where P_i is the actual normalized value, Q_i is the predicted value of the SVM model respectively and n is the number of data sets (days). The kernel function with smallest error measures is identified as the best kernel function of the SVM model. The smaller the values of error measures, the more accurate the model to be used in forecasting.

Table 2 represents the values of MAE and RMSE of these four kernel function for the air pollution index in Kota Bharu and Tanah Merah. As highlighted** in Table 2, RBF is the best kernel function since it has the lowest values of error measures while Sigmoid function has shown poor performance with highest values of MAE and RMSE. For Linear and Polynomial, their value of MAE and RMSE seems to be close to each other.

Table 2: Error Measures of Four Different Types of Kernel Functions

Kota Bharu				
	LINEAR	POLYNOMIAL	RBF	SIGMOID
MAE	0.1054539	0.1049263	0.09937916**	1.13694
RMSE	0.1457677	0.1458072	0.135974**	1.287093
Tanah Merah				
MAE	0.0920343	0.09035476	0.08550793**	0.9857326
RMSE	0.1274873	0.1260434	0.1192647**	1.137912

iii. Parameter Selection

Since the best kernel function is already chosen which is Radial Basis Function (RBF), then further investigation is on the selection of the best parameters value; cost (C) and gamma (γ) which have been set in the listed range of 0.001-100. The selection of the best pair of parameters value with the lowest error measure has been done by using the 10-fold cross-validation sampling method. The

best pair of parameters will be set on the fitting of SVM model training, validation of SVM model testing and SVM model to forecast future value.

Figure 3 shows the performance of SVM by tuning the parameters using the *tune()* command that has been done in the R programming system for API in Kota Bharu. The darkest area is represents the best pair of parameters ; cost = 100 and gamma = 100 with the error measure below than 0.010 which is 0.009743741.

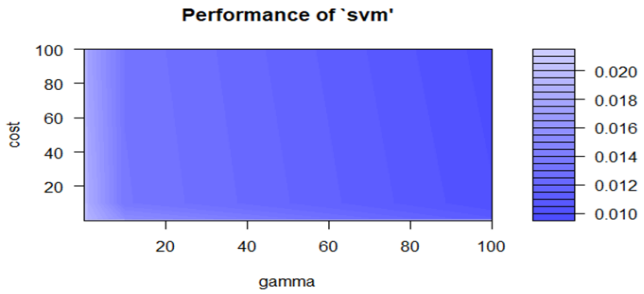

Figure 3: The Performance of Tuning SVM for API in Kota Bharu

iv. Fitting on Model Training

Fitting on model training has been done after obtaining the best kernel functions and its best parameters. The RBF kernel function for SVM model training is used by setting its best parameters which are cost = 100 and gamma = 100. If the result from this model training is not satisfactory, the process of parameter selection will be iterated by changing the appropriate range value of the parameters.

Tables 3 represents the values of MAE and RMSE on model training for API in Kota Bharu and Tanah Merah. As shown in the table, the values of MAE and RMSE with the best parameter is smaller than the default parameter. Since the values of error measures is decreasing, it shows that the forecasting process using the current SVM model is also more accurate rather than by setting default parameters.

Table 3: Error Measures on Model Training for API in Kota Bharu and Tanah Merah

	Kota Bharu		Tanah Merah	
	Default Parameter	Best Parameter	Default Parameter	Best Parameter
MAE	0.09937916	0.05045014	0.08550793	0.04186376
RMSE	0.135974	0.07651182	0.1192647	0.06518822

v. Validation on Model Testing

Once the model training of SVM is satisfactory, then the SVM model testing is validated. Tables 4 indicates the value of MAE and RMSE on model testing and model training for API in Kota Bharu and Tanah Merah. The values of MAE and RMSE on testing data for both locations are smaller than the training data. This can be concluded that the SVM model with RBF kernel function and its best parameter gave the best performance in forecasting the future values of API. The lowest error measure shows that the predicted value on training and testing data is closed to the actual data.

Table 4: Error Measures on Training and Testing Data for API in Kota Bharu and Tanah Merah

	Kota Bharu		Tanah Merah	
	Training data	Testing data	Training data	Testing data
MAE	0.05045014	0.03868583	0.04186376	0.03857308
RMSE	0.07651182	0.06251793	0.06518822	0.05895648

III) Data Implementation

After all process of data development was satisfactory, then the forecasting system of SVM model with the selected kernel function and the appropriate parameter is constructed. In this study, API is forecasted for the proceedings 12 days.

FINDINGS AND DISCUSSION

The default parameter was set for all the kernel functions in the model training, the RBF gives an accurate result with smallest error measures. The best value of two parameters, cost and gamma that are used for this kernel function is 100 for both parameters. The error measures decreased after the best parameter was set for model training, which are 0.05045014 (MAE) and 0.07651182 (RMSE) for API Kota Bharu and 0.04186376 (MAE) and 0.06518822 (RMSE) for API Tanah Merah.

In both locations, results showed that the MAE and RMSE for model testing are smaller than model training, which are 0.03868583 (MAE) and 0.06251793 (RMSE) for API in Kota Bharu and 0.03857308 (MAE) and 0.05895648 (RMSE) for API in Tanah Merah. Thus, the SVM model with RBF kernel function where both parameters are set to 100 gave a better performance in forecasting future values. Finally, the SVM model with RBF kernel function and its best parameter, is used to forecast API for the proceedings 12 days.

The prediction values of API in Kota Bharu and Tanah Merah are shown in Table 5 and Table 6, respectively. The values have been forecasted for 6 days in 2018 and 6 days in 2019 for API in Kota Bharu and 12 days in 2018 for API in Tanah Merah which are from day 345 until 356.

Table 5: The Prediction Value of API in Kota Bharu

Day	Prediction API KB
360	0.278063884
361	0.212891433
362	0.166075107
363	0.13792407
364	0.126922617
365	0.130399446
366	0.145168937
367	0.168043165
368	0.196168981
369	0.227193399
370	0.25929345
371	0.291120719

Table 6: The Prediction Value of API in Tanah Merah

Day	Prediction API TM
345	0.523349791
346	0.559828138
347	0.612018566
348	0.672264078
349	0.732283254
350	0.784636351
351	0.823800982
352	0.846724993
353	0.85285674
354	0.843757171
355	0.82245633
356	0.79272726

Figure 4 and Figure 5 show the predicted data for API in Kota Bharu and Tanah Merah, respectively.

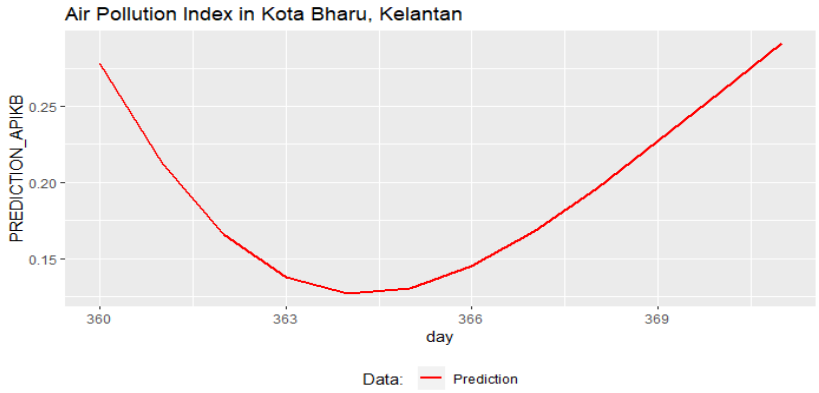

Figure 4: Graph of Predicted Data for API in Kota Bharu

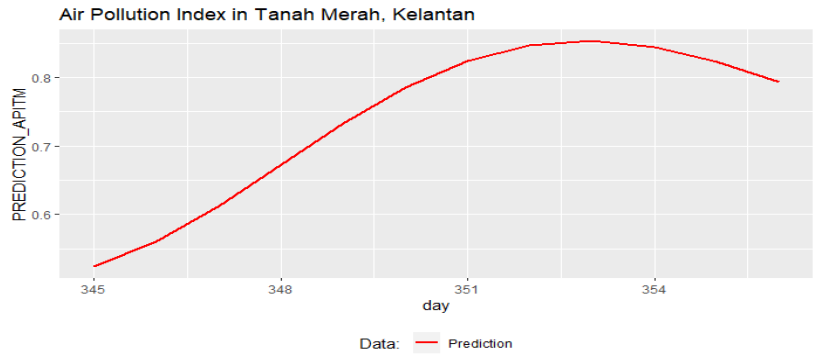

Figure 5: Graph of Predicted Data for API in Tanah Merah

Figure 4 showed the predicted API value in Kota Bharu decreases from day 360 to day 367 and then it increases until day 371. On the contrary, as shown in Table 6 and Figure 5, the predicted value for API in Tanah Merah increases for the first 9 days before starting to decrease from day 354 until day 356.

Figure 6 and Figure 7 illustrate the fluctuation of the data which are actual normalized, training, testing and predicted data for air pollution index in Kota Bharu and Tanah Merah, respectively.

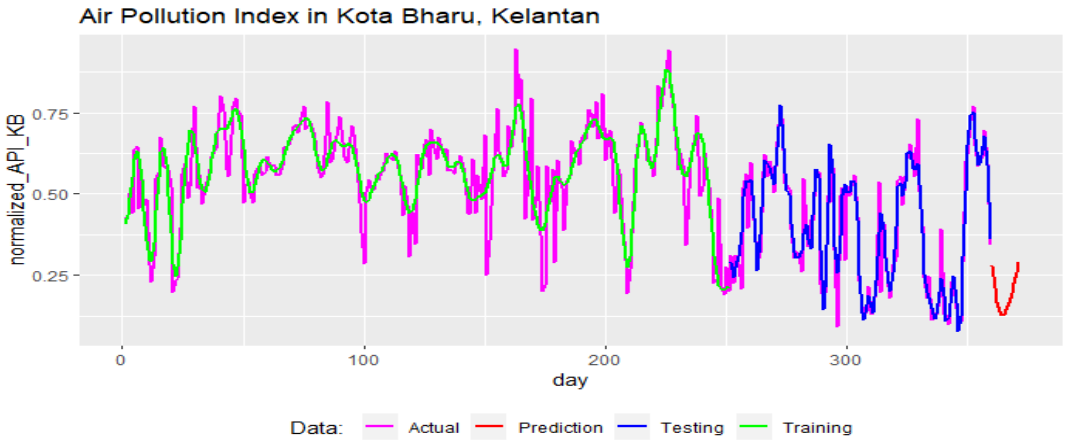

Figure 6: Graph of Actual Normalized, Training, Testing and Predicted Data for API in Kota Bharu

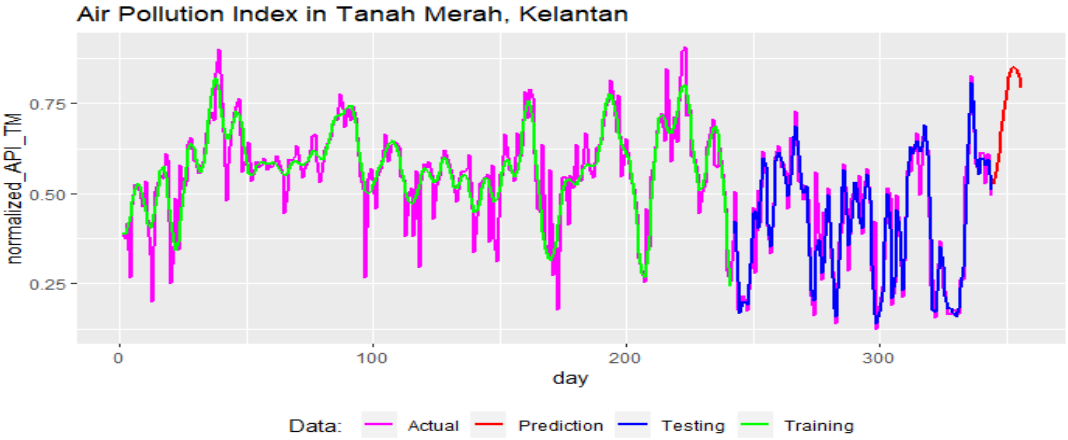

Figure 7: Graph of Actual Normalized, Training, Testing and Predicted Data for API in Tanah Merah

The horizontal axis of the graph displays the timeframe of day by day period, and the vertical axis represents the value of air pollution index. The forecasted value of API imitates the actual normalized data closely, which indicates good prediction.

CONCLUSION AND RECOMMENDATION

The main objective of this study is to access air pollution index of $PM_{2.5}$ using SVM and to compare the accuracy of four different types of the kernel function in SVM. The results showed that SVM works

relatively well and efficient in accessing air pollution index and the most appropriate kernel function of the SVM model to forecast the API with the smallest error measures is RBF kernel function.

This study has been done with only one method which is Support Vector Machine (SVM). In future it might be possible to analyze the same data by using hybrid method which is a combination of two or three model approaches such as SVM model with Artificial Neural Network (ANN) and Fuzzy Time Series (FTS). Furthermore, additional data and variables, for example other types of air pollutions, can be added for better understanding of the study.

ACKNOWLEDGMENTS

The authors sincerely thank all those who assisted in this research project. Their contributions were invaluable in completing the study.

CONFLICT OF INTERESTS DECLARATION

The authors declare no conflict of interests regarding the publication of this article.

REFERENCES

Arampongsanuwat, S., & Meesad, P. (2011). Prediction of PM10 using Support Vector Regression. *International Conference on Information and Electronics Engineering, IACSIT Press. Singapore.*

Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental pollution,* 151(2), 362-367.

Lu, W.-Z., & Wang, W.-J. (2005). Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. *Chemosphere,* 59(5), 693-701.

MAMPU, M. s. O. D. P. (2019). *Bacaan Indeks Pencemar Udara (IPU) bagi negeri Kelantan tahun 2018*. Retrieved Nov 01, 2019 from http://www.data.gov.my/data/en_US/dataset/bacaan-indeks-pencemar-udara-ipu-negeri-kelantan-bagi-tahun-2017/resource/25bbf752-661b-4445-8959-2ef45eaf1dfe

Oloruntoba, S., & Akinode, J. (2017). Student academic performance prediction using support vector machine. *International Journal of Engineering Sciences and Research Technology,* 6(12), 588-597.

Vaiz, J. S., & Ramaswami, M. (2016). A Hybrid Model to Forecast Stock Trend Using Support Vector Machine and Neural Networks.

Vinagre, E., Pinto, T., Ramos, S., Vale, Z., & Corchado, J. M. (2016). *Electrical energy consumption forecast using support vector machines. 27th International Workshop on Database and Expert Systems Applications (DEXA)*, 2016.